Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models

Year: 2024 | *Citations: 103* | *Authors: Wenhao Shi, Zhiqiang Hu, Yi Bin*

Abstract

Large language models (LLMs) have demonstrated impressive reasoning capabilities, particularly in textual mathematical problem-solving. However, existing open-source image instruction fine-tuning datasets, containing limited question-answer pairs per image, do not fully exploit visual information to enhance the multimodal mathematical reasoning capabilities of Multimodal LLMs (MLLMs). To bridge this gap, we address the lack of high-quality, diverse multimodal mathematical datasets by collecting 40K high-quality images with question-answer pairs from 24 existing datasets and synthesizing 320K new pairs, creating the MathV360K dataset, which enhances both the breadth and depth of multimodal mathematical questions. We introduce Math-LLaVA, a LLaVA-1.5-based model fine-tuned with MathV360K. This novel approach significantly improves the multimodal mathematical reasoning capabilities of LLaVA-1.5, achieving a 19-point increase and comparable performance to GPT-4V on MathVista's minitest split, and yielding leading performance on Math-V and MathVerse. Furthermore, Math-LLaVA demonstrates enhanced generalizability, showing substantial improvements on the MMMU benchmark. Our research highlights the importance of dataset diversity and synthesis in advancing MLLMs' mathematical reasoning abilities. The code and data are available at: \url{https://github.com/HZQ950419/Math-LLaVA}.